Recommended Statistical and Graphical Methods to Identify Potential Patterns of Anomalies Representing Suspicious Activities in Data Sets Collected from the Hunters Point Naval Shipyard Site

The identification of anomalous observations potentially representing falsification and suspicious activities in huge data sets collected from the various parcels of the Hunters Point Naval Shipyard Site (Site) over the past decade for many radionuclides of concern (ROCs) is a complex task. The use of advanced statistical and graphical methods especially designed to identify patterns present in complex multidimensional (for many ROCs) data sets is required for successful identification of anomalies and patterns potentially present in such data sets.

EPA recommends that Navy considers using effective univariate and multivariate/multidimensional statistical and graphical methods to identify potential suspicious/anomalous patterns present in data sets collected from the various parcels of the Site. There is no substitute for graphical displays generated using multivariate methods to identify potential patterns present a data set. Effective graphical methods provide added insight into the patterns present in a data set which is not possible to identify and understand simply based upon test statistics (e.g., K-S test statistic etc.). Once anomalous patterns have been identified using graphical displays, one can use statistical methods (e.g., hypothesis tests) to verify the existence of those patterns exhibited by the graphical displays.

For identified ROCs, EPA recommends the use multivariate methods which are better suited to effectively identify/recognize patterns present in a data set. Several ROCs are correlated (e.g., parent and daughter products), therefore multivariate methods which take correlations into consideration should be used. The use of such methods will considerably improve the likelihood of finding patterns and signs of falsification in a straight forward manner. The principal component analysis (PCA), factor analysis and classification analysis are commonly used to identify patterns in multidimensional data sets. These methods are meant to effectively identify patterns simultaneously for multiple variables (ROCs) included in the data set.

The recommended approaches described above have been used on the North Pier data sets. The effectiveness of the recommended multivariate methods has been illustrated using PCA on survey unit 1 (U1) and survey unit 7 (U7) multidimensional data sets. PCA has been performed on multivariate data set based upon ROCs: Cs-137, Bi-212, Pb-212, Bi-214, Pb-214, Ra-226- Bi 214, and Th 232/AC-228. Additional evaluations for these two survey units are provided in Appendix A.

Evaluation of the North Pier Parcel Data Using Multivariate PCA Method for U1 and U7

At the North Pier parcel, only two rounds of systematic sampling: Sys-1 and Sys-2 were performed. Typically, observed values of a ROC are the highest during the first round of sampling Sys-1. Overall, values of ROCs should be the lowest during the final status survey, FSS-Sys and the highest during Sys-1. ROC values observed during Sys-2 phase should also be higher than FSS-Sys (after two rounds of sampling and remediation). If this "desired" pattern is not followed by observed values of ROCs during sampling phases, it may be inferred that data have been manipulated/falsified.

Using the recommended methods, suspicious patterns have been identified for all ROCs (included in the evaluation) during sampling phase (s) and collection dates. At the North Pier site, there are two other sampling phases: Biased FSS and RAS. Statistical methods have been used on all 5 sampling phases: Sys-1, Sys-2, FSS-Sys, Bias-Sys, and RAS but comments have been provided based upon the comparison of Sys-1, Sys-2, and FSS-Sys data.

Scatter Plots of the First Two Principal Components (PCs)-U1: The first two PCs account for the majority of information (on all ROCs) present in a multivariate data set. For U1, the first two PCs explain about 88% of information (Figures 1 and 2) present in the multidimensional data set. Based upon data from U1, Figure 1 has graphical display of the first two principal components: PC1 and PC2 by sampling phases and Figure 2 has graphical displays of the first two PCs by collection dates.

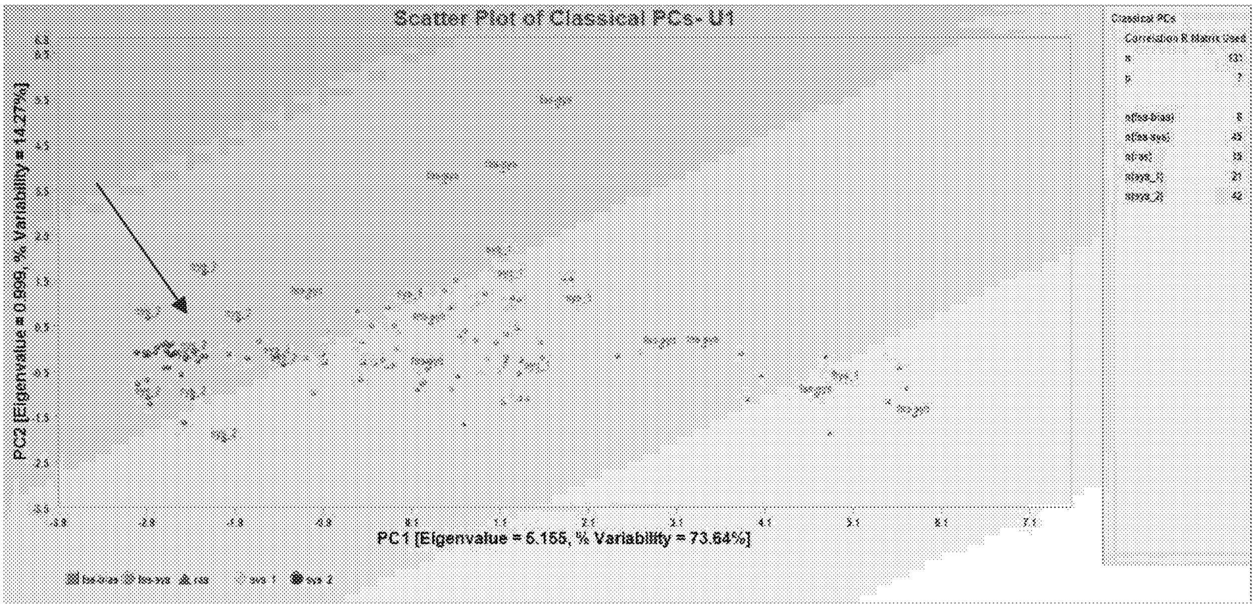


Figure 1. Scatter Plot of PC1 versus PC2 by Sampling Phases – Survey Unit 1

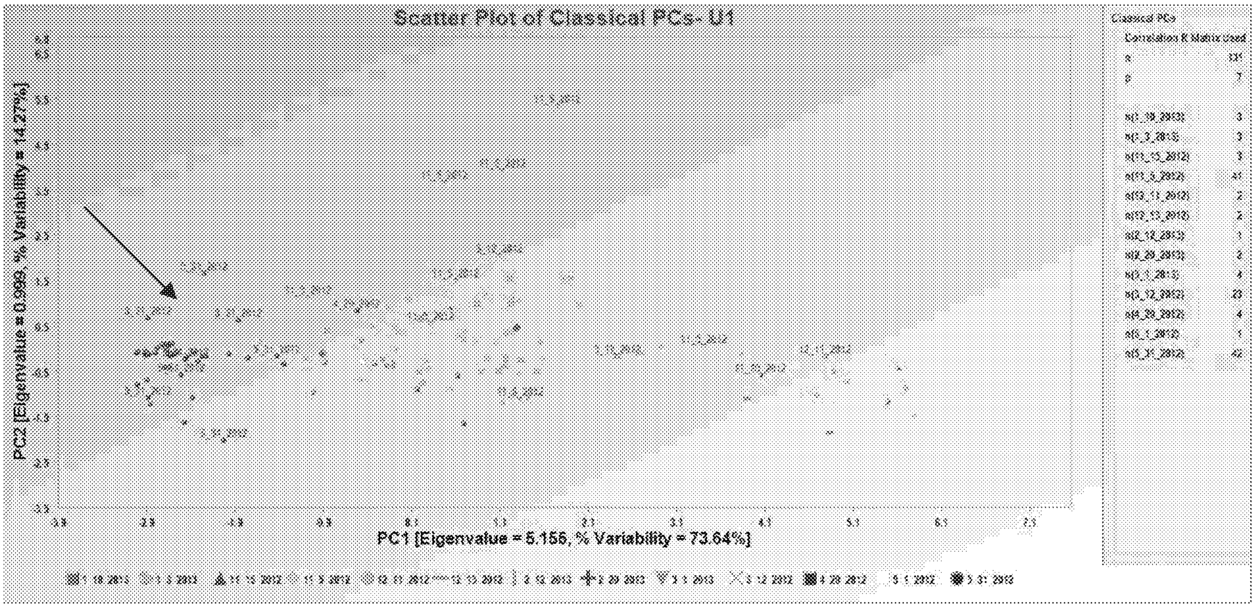
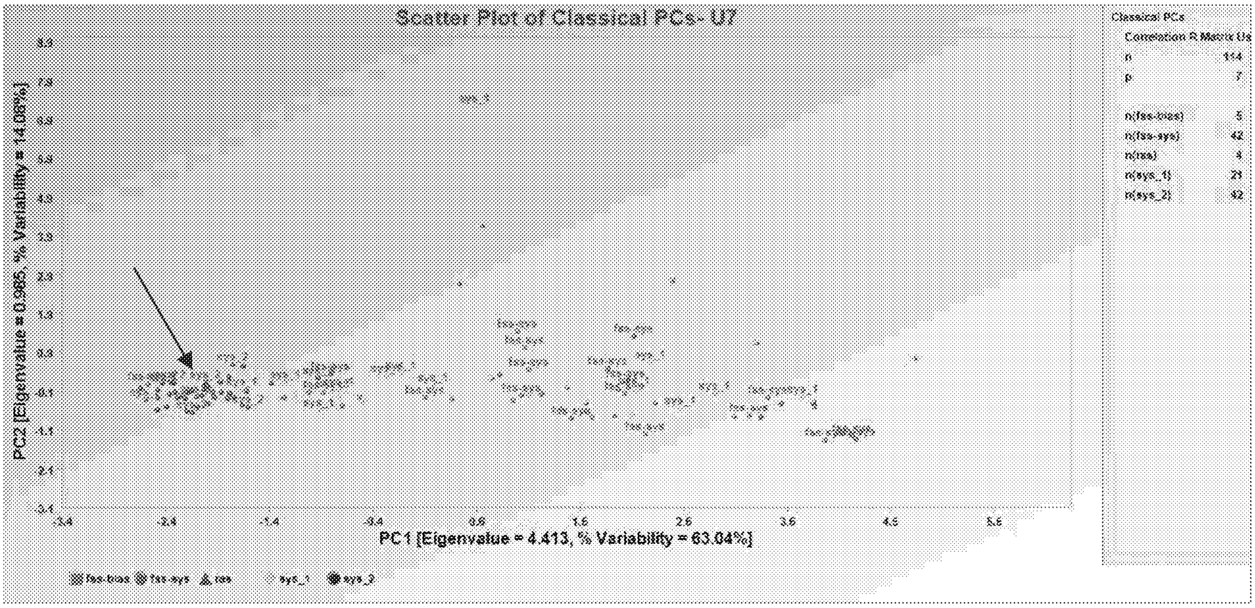


Figure 2. Scatter Plot of PC1 versus PC2 by Collection Dates – Survey Unit 1

Survey Unit1: From pattern displayed in Figure 1, it is noted that Sys-2 data set (42 observations) is tightly clustered (identified by a red arrow) with reduced variability and is well separated from the rest of the data. This pattern leads to the conclusion that some different (suspicious) activities might have taken place during Sys-2 phase. Similarly, from Figure 2 it is noted that data collected on May 31, 2012 (42 observations) is tightly clustered (identified by a red arrow) and is well separated from the rest of the data. This pattern leads to the conclusion that some different (suspicious) activities might have taken place during the sample collection performed on May 31, 2012.

- *These two graphs alone identified potentially suspicious/altered data in U1 collected on May 31, 2012 during sampling phase Sys-2 for all ROCs considered in PCA evaluations.*

Scatter Plots of the First Two Principal Components (PCs)-U7: The first two PCs based upon U7 data set explain about 77% of the information (Figures 3 and 4) contained in the multidimensional (for all ROCs considered) data set. Based upon data from U7, Figure 3 has graphical display of the first two principal components: PC1 and PC2 by sampling phases and Figure 4 has graphical displays of the first two PCs by collection dates.

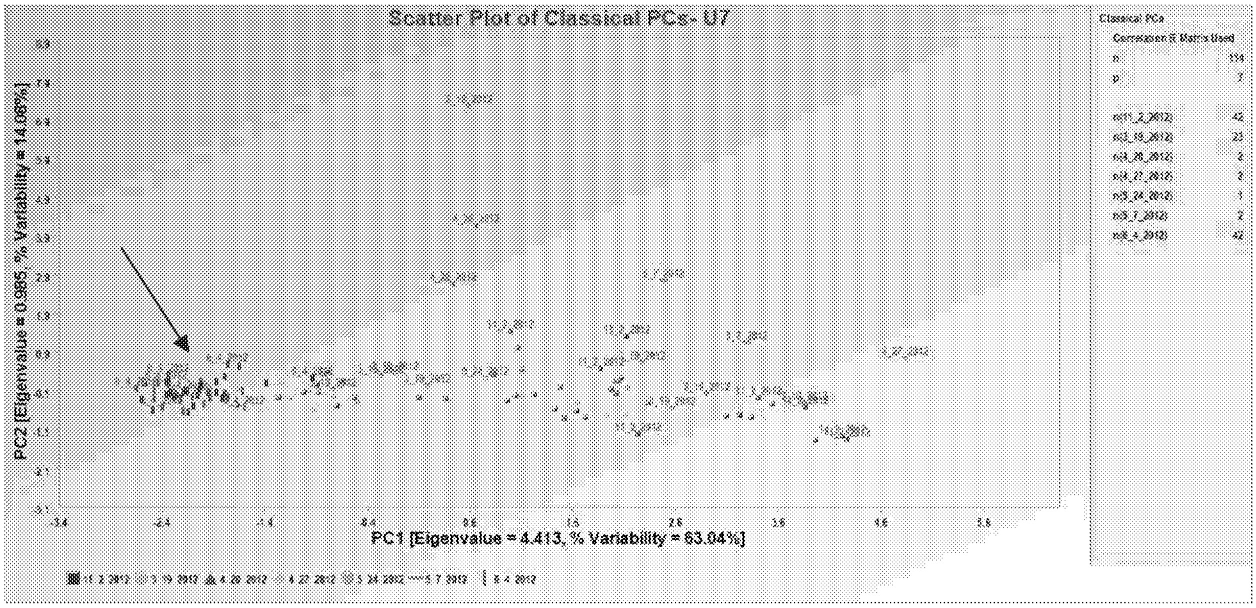


Classical PCs

Correlation R Matrix Us

n	114
p	7
n(fss-bias)	5
n(fss-sys)	42
n(rss)	4
n(sys_1)	21
n(sys_2)	42

Figure 3. Scatter Plot of PC1 versus PC2 by Sampling Phases – Survey Unit 7



Classical PCs

Correlation R Matrix Used

n	114
p	7
n(11_2_2012)	42
n(3_19_2012)	23
n(4_26_2012)	2
n(4_27_2012)	2
n(5_24_2012)	1
n(5_7_2012)	2
n(6_4_2012)	42

Figure 4. Scatter Plot of PC1 versus PC2 by Collection Dates – Survey Unit 7

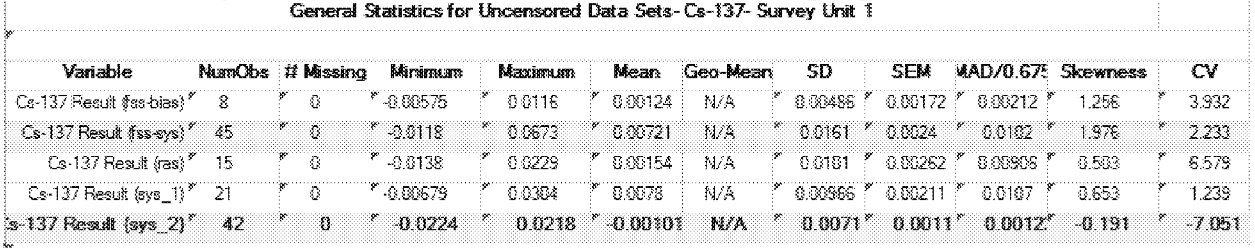
Survey Unit 7: From pattern displayed in Figure 3, it is noted that Sys-2 data (42 observations) is tightly clustered (pointed by a red arrow) with reduced variability and is well separated from the rest of the data. This pattern leads to the conclusion that some different (suspicious) activities might have taken place during Sys-2 phase. Similarly, in Figure 12 it is noted that data collected on June 4, 2012 (42 observations) is tightly clustered (pointed by a red arrow) and is well separated from the rest of the data. This pattern leads to the conclusion that some different (suspicious) activities might have taken place during the sample collection performed on June 4, 2012.

- *These two graphs (Figures 3 and 4) alone identified potentially suspicious/altered data in U7 collected on June 4, 2012 during sampling phase Sys-2 for all ROCs simultaneously.*

Looking Deeper in Survey Unit 1 and 7 Data Sets

One may want to look deeper into data sets collected from U1 and U7 to determine what happened on May 31, 2012 in U1 and on June 4, 2012 in U7. Summary Statistics were computed. Table 1 has summary statistics for Cs-137 by sampling phases in U1 and Table 2 has summary statistics for Cs-137 by sample collection dates for U1.

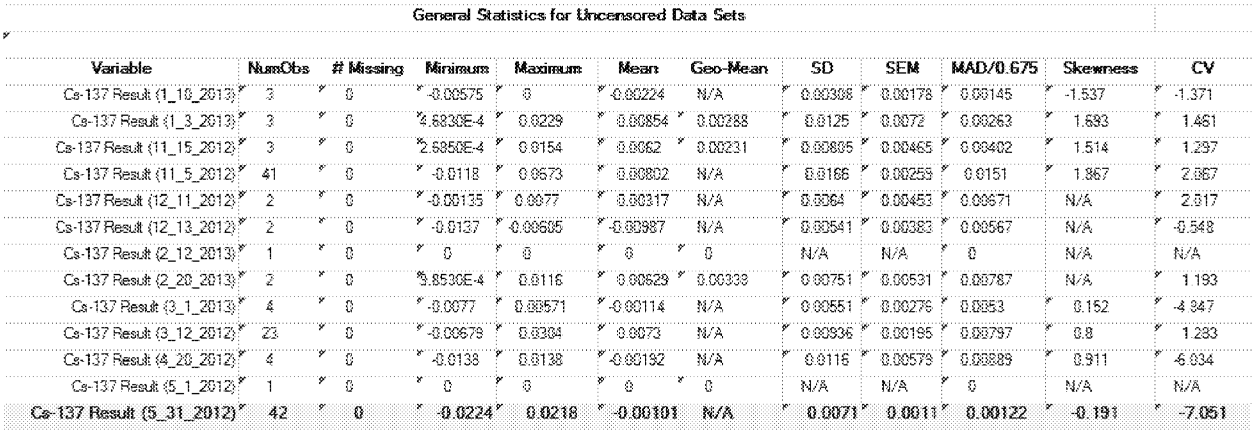
Table 1. Summary Statistics for Cs-137 in Survey Unit 1 by Sampling Phases



General Statistics for Uncensored Data Sets- Cs-137- Survey Unit 1

Variable	NumObs	# Missing	Minimum	Maximum	Mean	Geo-Mean	SD	SEM	MAD/0.675	Skewness	CV
Cs-137 Result (fss-bias)	8	0	-0.00575	0.0116	0.00124	N/A	0.00486	0.00172	0.00212	1.256	3.932
Cs-137 Result (fss-sys)	45	0	-0.0118	0.0673	0.00721	N/A	0.0161	0.0024	0.0102	1.976	2.233
Cs-137 Result (ras)	15	0	-0.0138	0.0229	0.00154	N/A	0.0101	0.00262	0.00906	0.503	6.579
Cs-137 Result (sys_1)	21	0	-0.00679	0.0304	0.0078	N/A	0.00966	0.00211	0.0107	0.653	1.239
s-137 Result (sys_2)	42	0	-0.0224	0.0218	-0.00101	N/A	0.0071	0.0011	0.0012	-0.191	-7.051

Table 2. Summary Statistics for Cs-137 in Survey Unit 1 by Collection Dates



General Statistics for Uncensored Data Sets

Variable	NumObs	# Missing	Minimum	Maximum	Mean	Geo-Mean	SD	SEM	MAD/0.675	Skewness	CV
Cs-137 Result (1_10_2013)	3	0	-0.00575	0	-0.00224	N/A	0.00308	0.00178	0.00145	-1.537	-1.371
Cs-137 Result (1_3_2013)	3	0	4.6830E-4	0.0229	0.00854	0.00288	0.0125	0.0072	0.00263	1.693	1.461
Cs-137 Result (11_15_2012)	3	0	2.6850E-4	0.0154	0.0062	0.00231	0.00805	0.00465	0.00402	1.514	1.297
Cs-137 Result (11_5_2012)	41	0	-0.0118	0.0673	0.00802	N/A	0.0166	0.00259	0.0151	1.867	2.067
Cs-137 Result (12_11_2012)	2	0	-0.00135	0.0077	0.00317	N/A	0.0064	0.00453	0.00671	N/A	2.017
Cs-137 Result (12_13_2012)	2	0	-0.0137	-0.00605	-0.00987	N/A	0.00541	0.00383	0.00567	N/A	-0.548
Cs-137 Result (2_12_2013)	1	0	0	0	0	0	N/A	N/A	0	N/A	N/A
Cs-137 Result (2_20_2013)	2	0	9.8530E-4	0.0116	0.00629	0.00338	0.00751	0.00531	0.00787	N/A	1.193
Cs-137 Result (3_1_2013)	4	0	-0.0077	0.00571	-0.00114	N/A	0.00551	0.00276	0.0053	0.152	-4.847
Cs-137 Result (3_12_2012)	23	0	-0.00679	0.0304	0.0073	N/A	0.00936	0.00195	0.00797	0.8	1.283
Cs-137 Result (4_20_2012)	4	0	-0.0138	0.0138	-0.00192	N/A	0.0116	0.00579	0.00889	0.911	-6.034
Cs-137 Result (5_1_2012)	1	0	0	0	0	0	N/A	N/A	0	N/A	N/A
Cs-137 Result (5_31_2012)	42	0	-0.0224	0.0218	-0.00101	N/A	0.0071	0.0011	0.00122	-0.191	-7.051

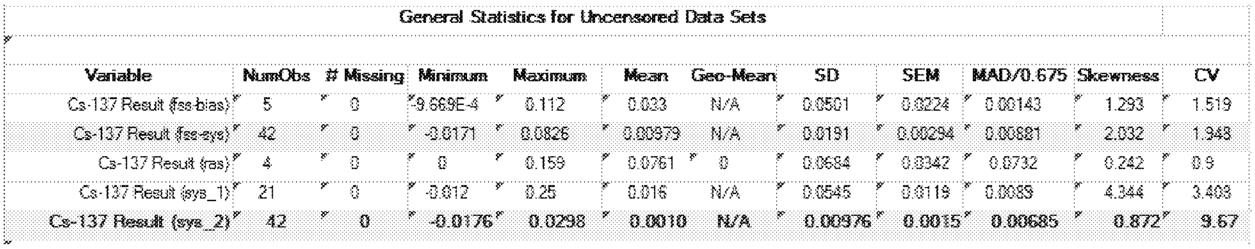
- Note that for survey unit 1, on May 31, 2012, 42 samples were evaluated during phase Sys-2.

- Table 1: Note data for Sys-2 (out of 3 phases) phase exhibits the lowest mean, lowest value of the maximum value, and the lowest standard deviation (sd). These values might have been manipulated during this phase to reduce mean and variability (explaining tight clustering for Sys-2 as shown in Figure 1).

- Table 2: Data for collection date May 31, 2012 exhibits the lowest mean, lowest value of the maximum value, and lowest standard deviation (sd) among all dates with more than 4 samples. Cs-137 values on this date might have been manipulated to reduce mean and variability (explaining tight clustering for this date shown in Figure 2).

Table 3 has summary statistics for Cs-137 by sampling phases in U7 and Table 4 has summary statistics for Cs-137 by sample collection dates for U7.

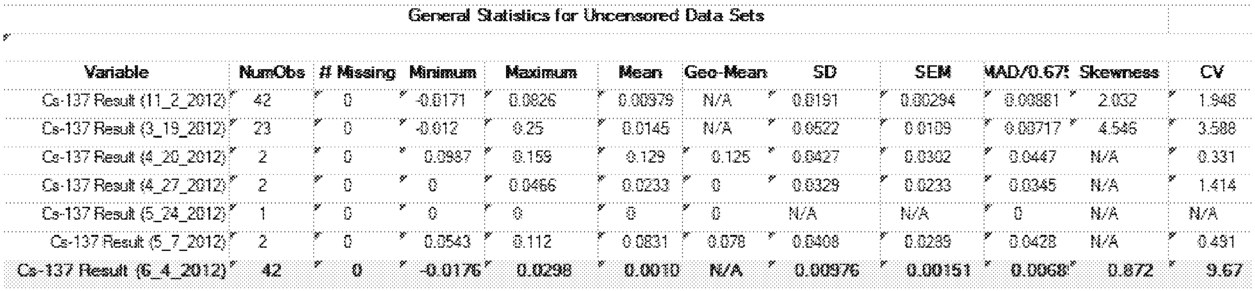
Table 3. Summary Statistics for Cs-137 in Survey Unit 7 by Sampling Phases



General Statistics for Uncensored Data Sets

Variable	NumObs	# Missing	Minimum	Maximum	Mean	Geo-Mean	SD	SEM	MAD/0.675	Skewness	CV
Cs-137 Result (fss-bias)	5	0	-9.669E-4	0.112	0.033	N/A	0.0501	0.0224	0.00143	1.293	1.519
Cs-137 Result (fss-sys)	42	0	-0.0171	0.0826	0.00979	N/A	0.0191	0.00294	0.00881	2.032	1.948
Cs-137 Result (ras)	4	0	0	0.159	0.0761	0	0.0684	0.0342	0.0732	0.242	0.9
Cs-137 Result (sys_1)	21	0	-0.012	0.25	0.016	N/A	0.0545	0.0119	0.0089	4.344	3.408
Cs-137 Result (sys_2)	**42**	**0**	**-0.0176**	**0.0298**	**0.0010**	**N/A**	**0.00976**	**0.0015**	**0.00685**	**0.872**	**9.67**

Table 4. Summary Statistics for Cs-137 in Survey Unit 7 by Sampling Phases



General Statistics for Uncensored Data Sets

Variable	NumObs	# Missing	Minimum	Maximum	Mean	Geo-Mean	SD	SEM	MAD/0.675	Skewness	CV
Cs-137 Result (11_2_2012)	42	0	-0.0171	0.0826	0.00979	N/A	0.0191	0.00294	0.00881	2.032	1.948
Cs-137 Result (3_19_2012)	23	0	-0.012	0.25	0.0145	N/A	0.0522	0.0109	0.00717	4.546	3.588
Cs-137 Result (4_20_2012)	2	0	0.0987	0.159	0.129	0.125	0.0427	0.0302	0.0447	N/A	0.331
Cs-137 Result (4_27_2012)	2	0	0	0.0466	0.0233	0	0.0329	0.0233	0.0345	N/A	1.414
Cs-137 Result (5_24_2012)	1	0	0	0	0	0	N/A	N/A	0	N/A	N/A
Cs-137 Result (5_7_2012)	2	0	0.0543	0.112	0.0831	0.078	0.0408	0.0289	0.0428	N/A	0.491
Cs-137 Result (6_4_2012)	**42**	**0**	**-0.0176**	**0.0298**	**0.0010**	**N/A**	**0.00976**	**0.00151**	**0.0068**	**0.872**	**9.67**

- Note that for survey unit 7, on June 4, 2012, 42 samples were evaluated during phase Sys-2.

- Table 3: Data for Sys-2 (out of 3 phases) phase exhibits the lowest mean, lowest value of the maximum value, and lowest sd. Values might have been manipulated during this phase to lower the mean and variability (explaining tight clustering for Sys-2 as shown in Figure 3).

- Table 4: Data for collection date June 4, 2012 exhibits the lowest mean, lowest value of the maximum value, and lowest sd among all dates with more than 4 samples. Values might have been manipulated on this date to reduce mean and variability (explaining tight clustering for June 4, 2012 as shown in Figure 4).

Summary: As demonstrated by patterns displayed in Figures 1 through 4, suspicious activities in U1 and U7 for ROCs included in the evaluations have been identified using only four PC graphs (Figures 1 through 4). These graphs identified that suspicious activities/ falsification for all ROCs included in the evaluations took place mainly during Sys-2 sampling phase. In U1, suspicious activity took place on May 31, 2012 and in U7, suspicious activity took place on June 4, 2012.

For Cs-137, these conclusions are supplemented with statistics displayed in Tables 1 through 4 for Cs-137. If deemed necessary, one may want to generate these statistics tables for all other ROCs. However, multivariate methods identified suspicious activities simultaneously for all ROCs included in the evaluations. Also, if deemed necessary, one can verify the conclusions derived based upon PC evaluations described above by using scatter plots of the first PC1 against ROCs considered in PC evaluations. Additionally, one can also use univariate graphical and statistical methods. These evaluations are summarized in Appendix A.